

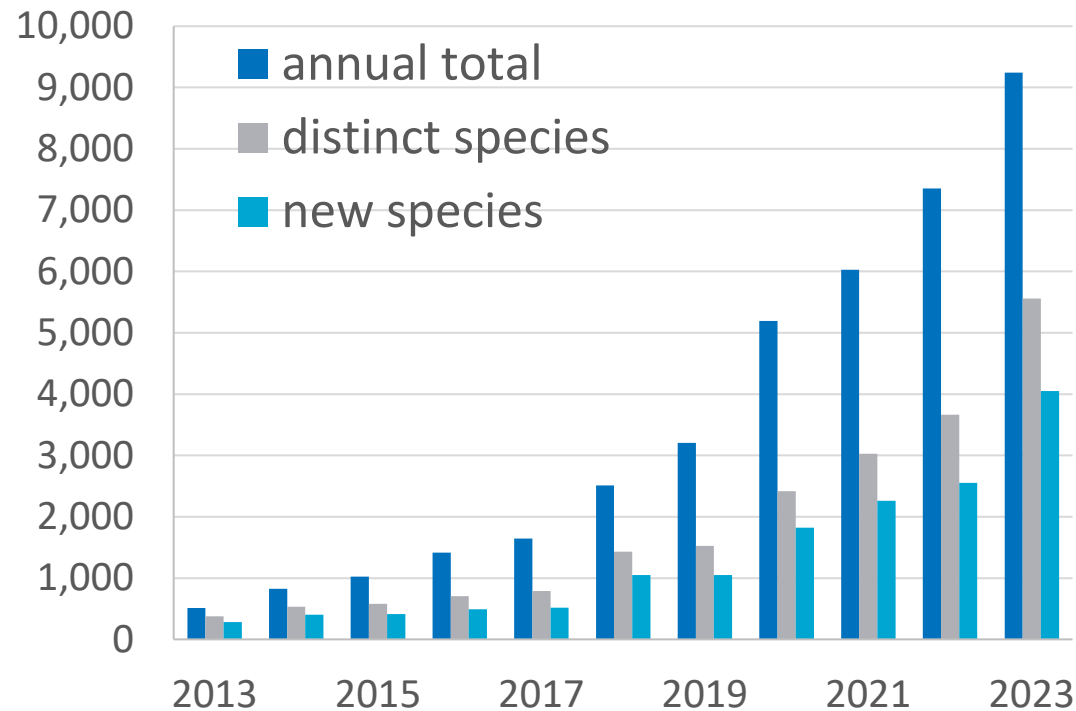
The NIH Comparative Genomics Resource (CGR)

Terence D. Murphy – CGR Project Lead

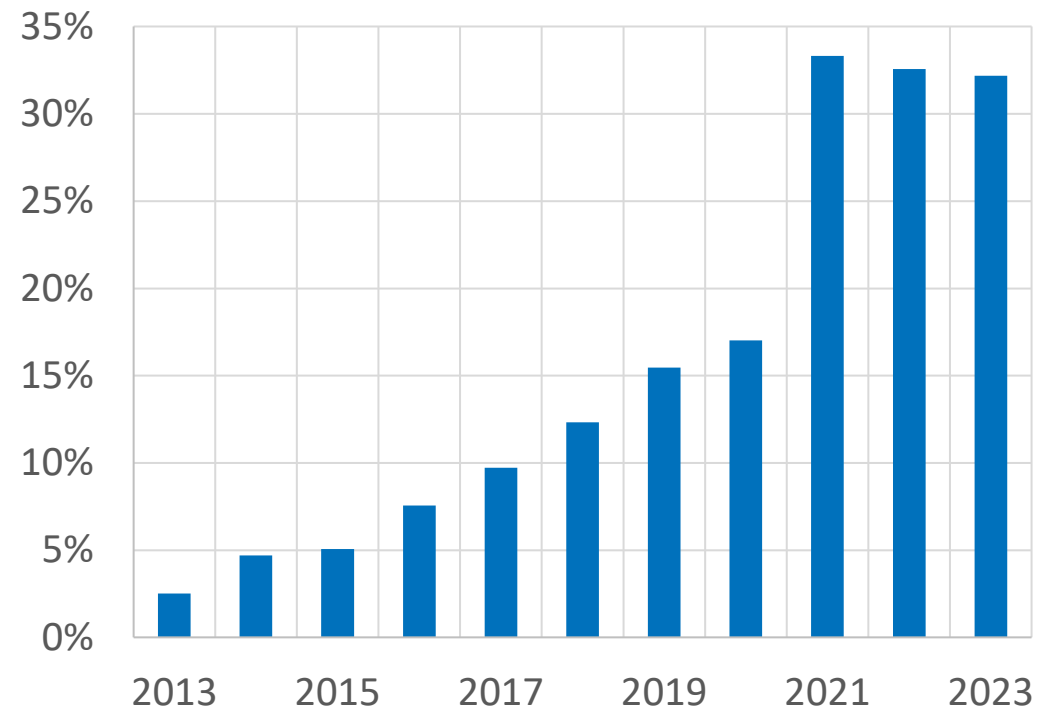
Look at all the genomes!

Current eukaryotic genomes
(as of Dec 31 2023):
36,593 genomes
15,453 species

Annual Eukaryote Genome Submissions



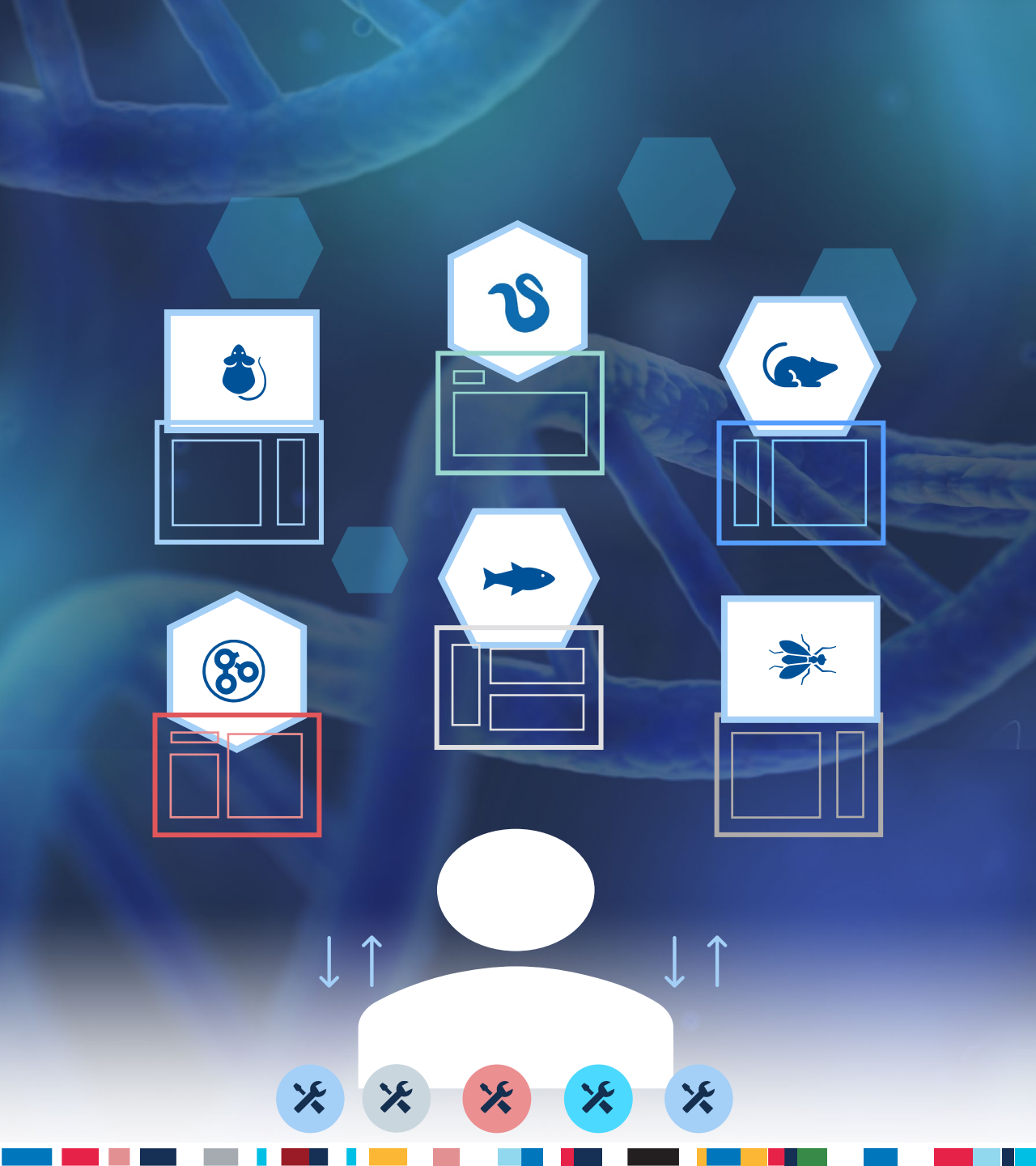
Percent with ContigN50 > 1 Mbp



Why CGR?

Comparative genomics investigators face several limitations and challenges

- Multiple different user interfaces
- Limited number of organisms supported
- Siloed data and applications
- Must download data to apply tools
- Limited scalability
- Data quality issues

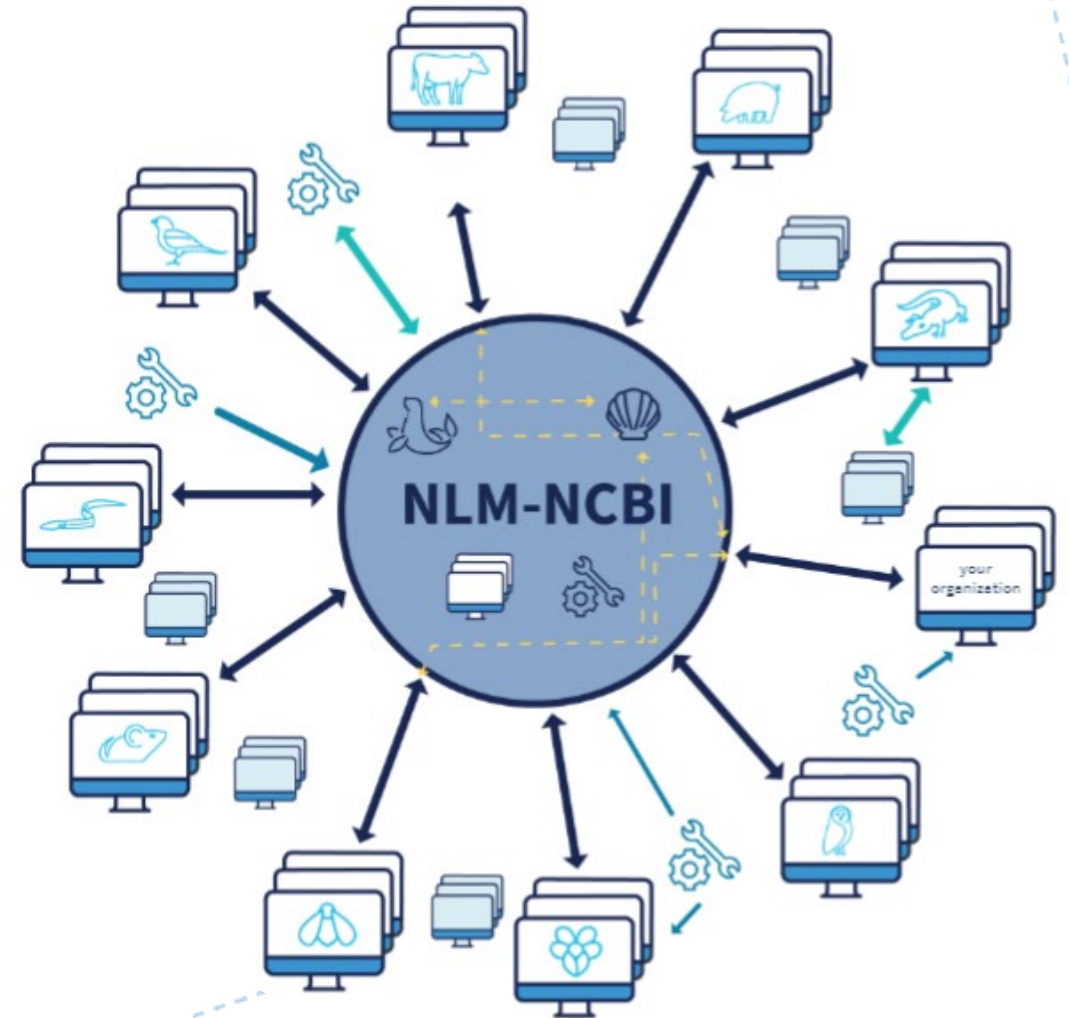


Solution

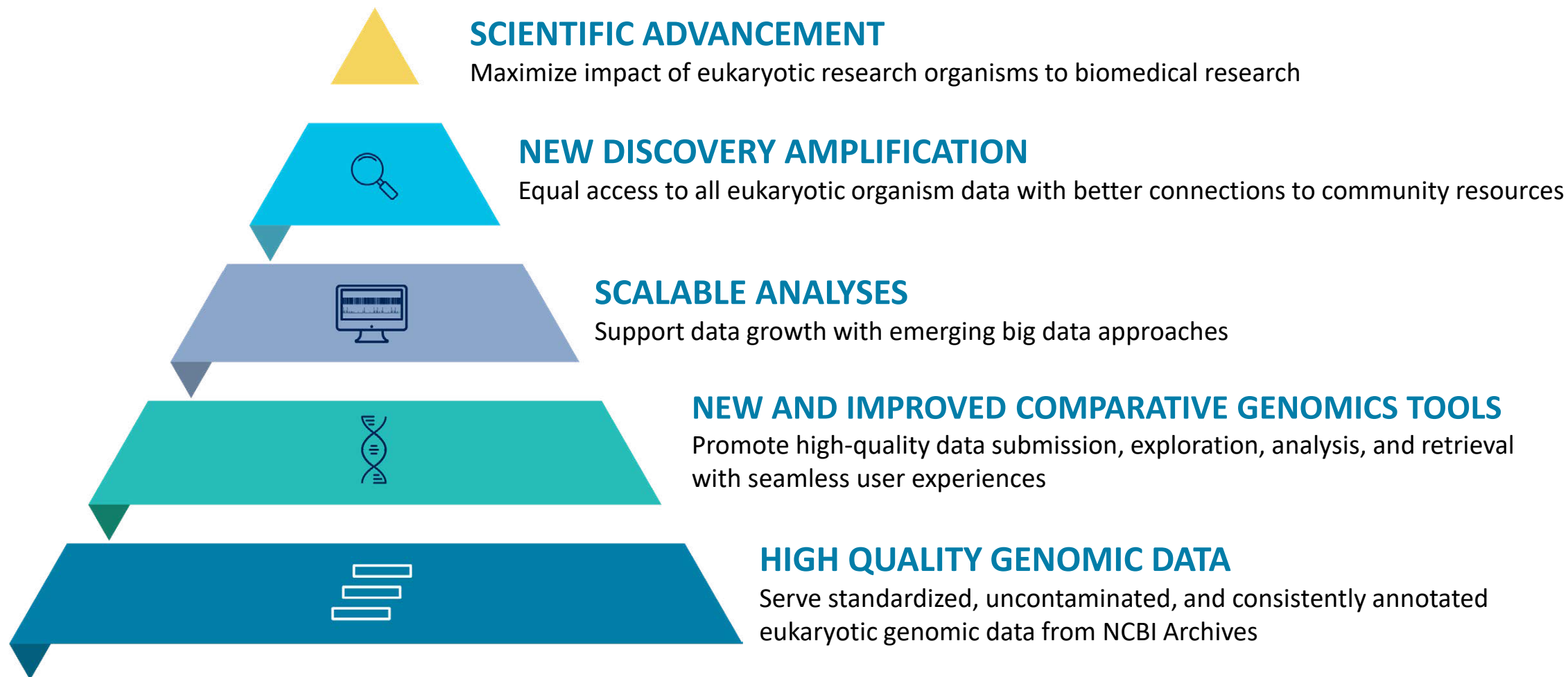
NIH Comparative Genomics Resource (CGR)

What: CGR **maximizes the impact of eukaryotic research organisms** and their genomic data to biomedical research.

How: CGR facilitates reliable comparative genomics analyses through **community collaboration** and an **NCBI genomics toolkit**. The toolkit includes high-quality data, tools, and **interfaces** for connecting community-provided resources with NCBI.

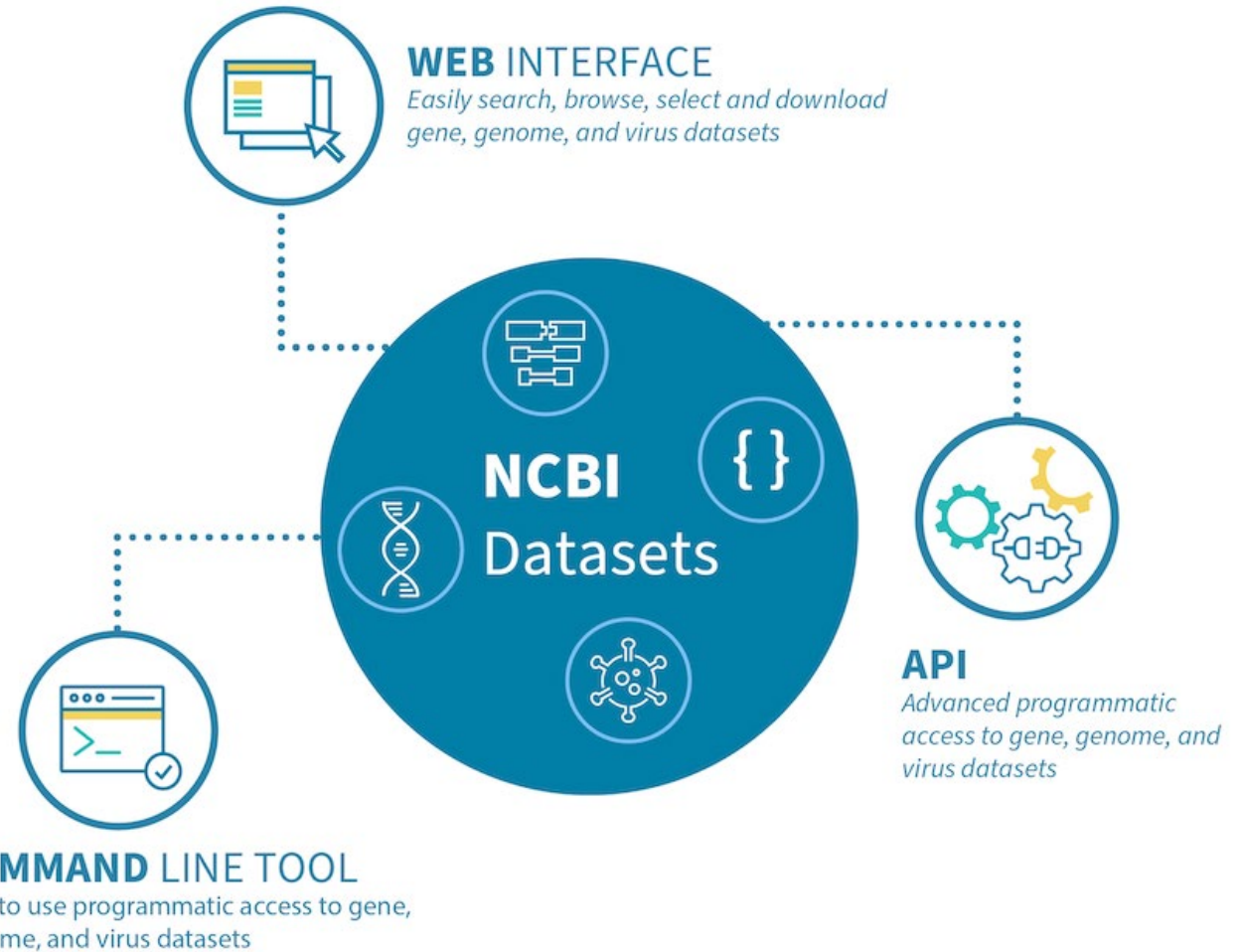


CGR Benefits



NCBI Datasets

- Easy-to-use web and programmatic interfaces
- Data delivery in formats that support both web and programmatic users
- Helpful documentation and tutorials



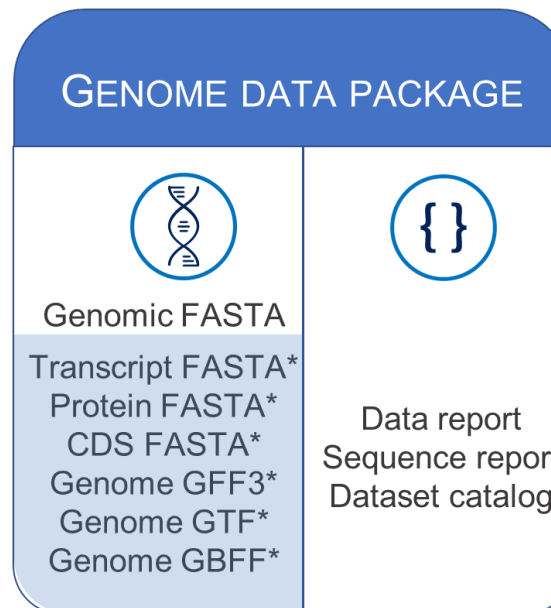


NCBI Datasets Genome Data Package

Data

- Genome FASTA
- Annotation (GFF, GTF, GBFF)
- Transcript, Protein, CDS FASTA

GENOME



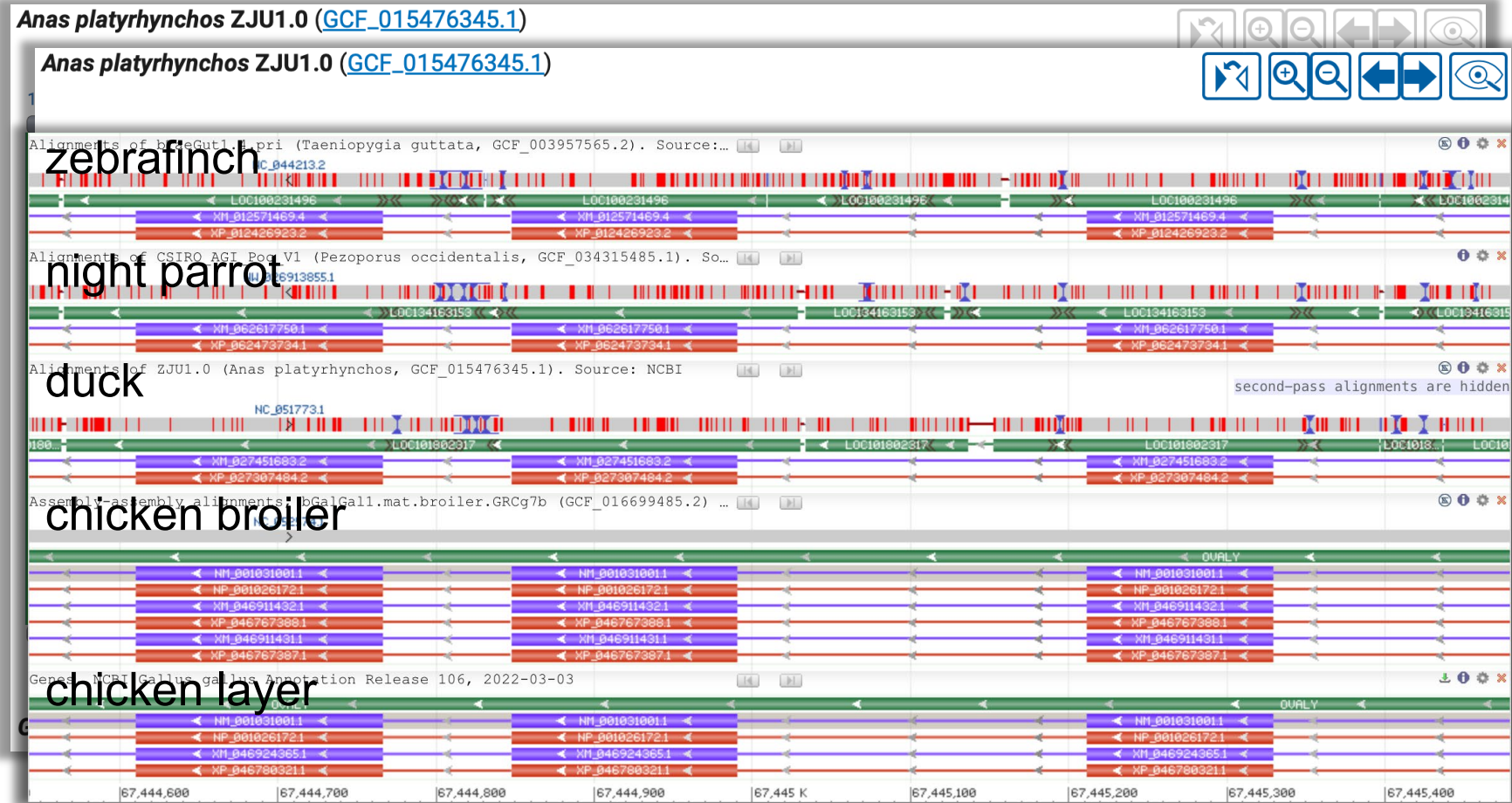
Metadata

- Contains metadata from Assembly, Gene, BioSample, and BioProject databases
- JSON lines format

Comparative Genome Viewer (CGV)

- Compare genomes through assembly alignments
- Over 700 alignments for over 300 species!
- Preprint: PMC10705539

<https://www.ncbi.nlm.nih.gov/genome/cgv/>



Clusters

Graphic Summary

NP_001358098.1 (placentals) - 199 member(s)

Reports

Cluster Members

Cluster Taxonomy

[BLAST Alignment](#)
[Tree View](#)
[MSA Viewer](#)

Showing 25 cluster member(s) out of 199

Accession	Scientific Name	Common Name	Taxid	Accession
NP_001358098.1	Homo sapiens	human	9606	N/A
AAC50765.1	Homo sapiens	human	9606	N/A
AAF22283.1	Homo sapiens	human	9606	N/A
CAA04507.1	Homo sapiens	human	9606	N/A
EAW83428.1	Homo sapiens	human	9606	N/A
KAF6086466.1	Phyllostomus discolor	pale spear-nosed bat	89673	N/A
KAF6319901.1	Myotis myotis	bats	51298	N/A
NP_001032209.1	Homo sapiens	human	9606	3 or more identical proteins
NP_001358067.1	Homo sapiens	human	9606	N/A
NP_001358074.1	Homo sapiens	human	9606	N/A
NP_001358085.1	Homo sapiens	human	9606	N/A
PNJ01110.1	Pan troglodytes	chimpanzee	9598	N/A
PNJ29883.1	Pongo abelii	Sumatran orangutan	9601	2 or more identical proteins
PNJ29892.1	Pongo abelii	Sumatran orangutan	9601	N/A
XP_002751796.3	Callithrix jacchus	white-tufted-ear marmoset	9483	2 or more identical proteins
XP_003268178.1	Nomascus leucogenys	northern white-cheeked gibbon	61853	N/A
XP_003921138.1	Saimiri boliviensis boliviensis	Bolivian squirrel monkey	39432	N/A

Download

Text
CSV
FASTA
GenBank

Show

100

?

[BLAST Alignment](#)
[MSA Viewer](#)

Percent Identity	Accession Length	Accession
100%	1292	NP_001358098.1
96%	1285	NP_001358053.1
95%	1288	NP_001358048.1
91%	1284	XP_008069675.1
86%	1207	XP_040858153.1
83%	1285	XP_033612730.1
83%	1282	XP_035866695.1
83%	1299	XP_014386430.1
84%	1301	XP_005335987.1
85%	1289	XP_033612726.1
82%	1210	XP_036353138.1
87%	1289	XP_017902699.1
82%	1243	XP_044991521.1
82%	1282	XP_012576331.1

select all

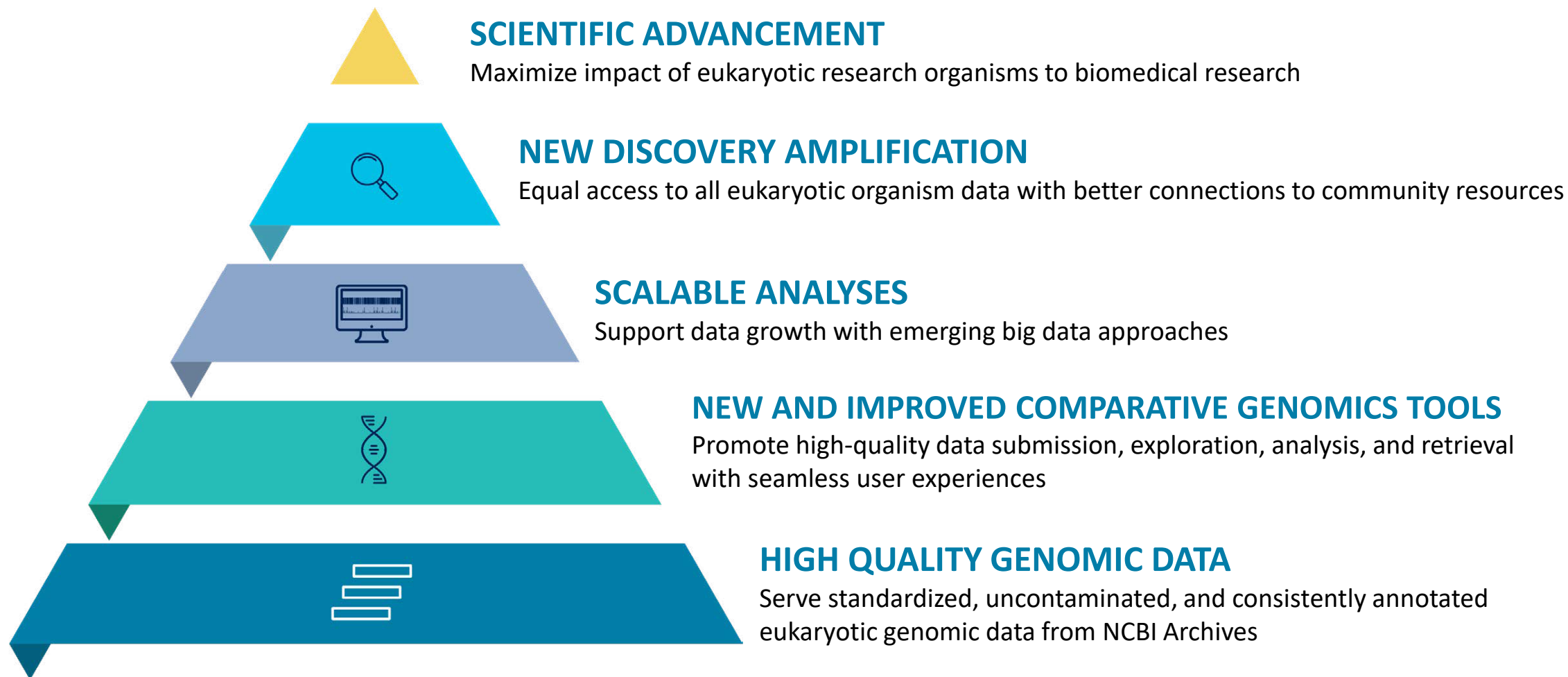
100 clusters selected

Cluster title

Click the icon to see the cluster content

☒	199 member(s), 39 organism(s)
☒	484 member(s), 75 organism(s)
☒	13 member(s), 8 organism(s)
☒	2 member(s), 1 organism(s)
☒	2 member(s), 2 organism(s)
☒	1 member(s), 1 organism(s)
☒	12 member(s), 4 organism(s)
☒	12 member(s), 4 organism(s)
☒	1 member(s), 1 organism(s)
☒	4 member(s), 4 organism(s)
☒	6 member(s), 2 organism(s)
☒	2 member(s), 2 organism(s)
☒	7 member(s), 3 organism(s)
☒	12 member(s), 1 organism(s)

CGR Benefits



Foreign Contaminant Screening (FCS)

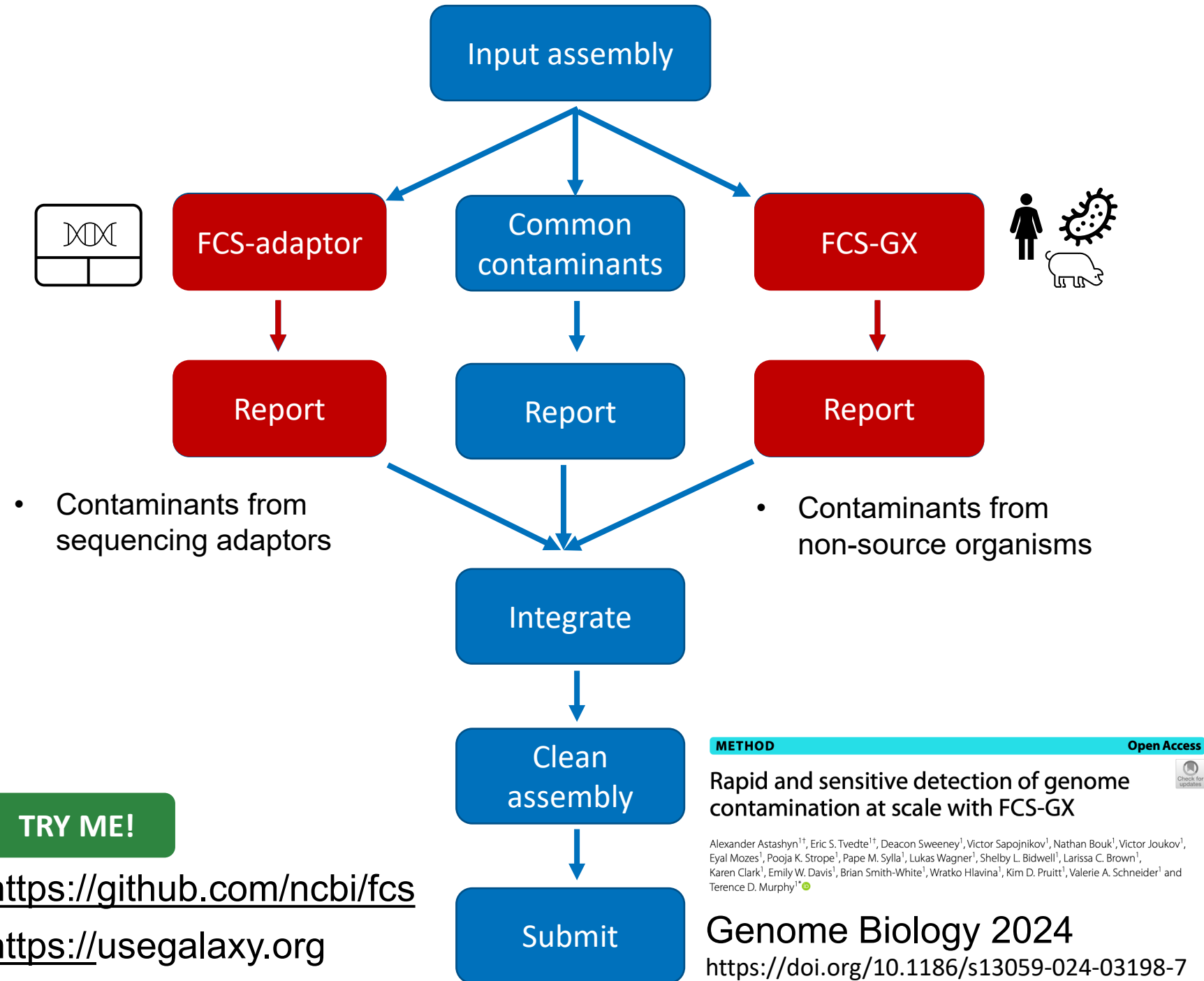
CGR Goals:

- Modular
- Easy
- Reliable
- High sensitivity and specificity
- Fast
- Same code used for GenBank submission screening

TRY ME!

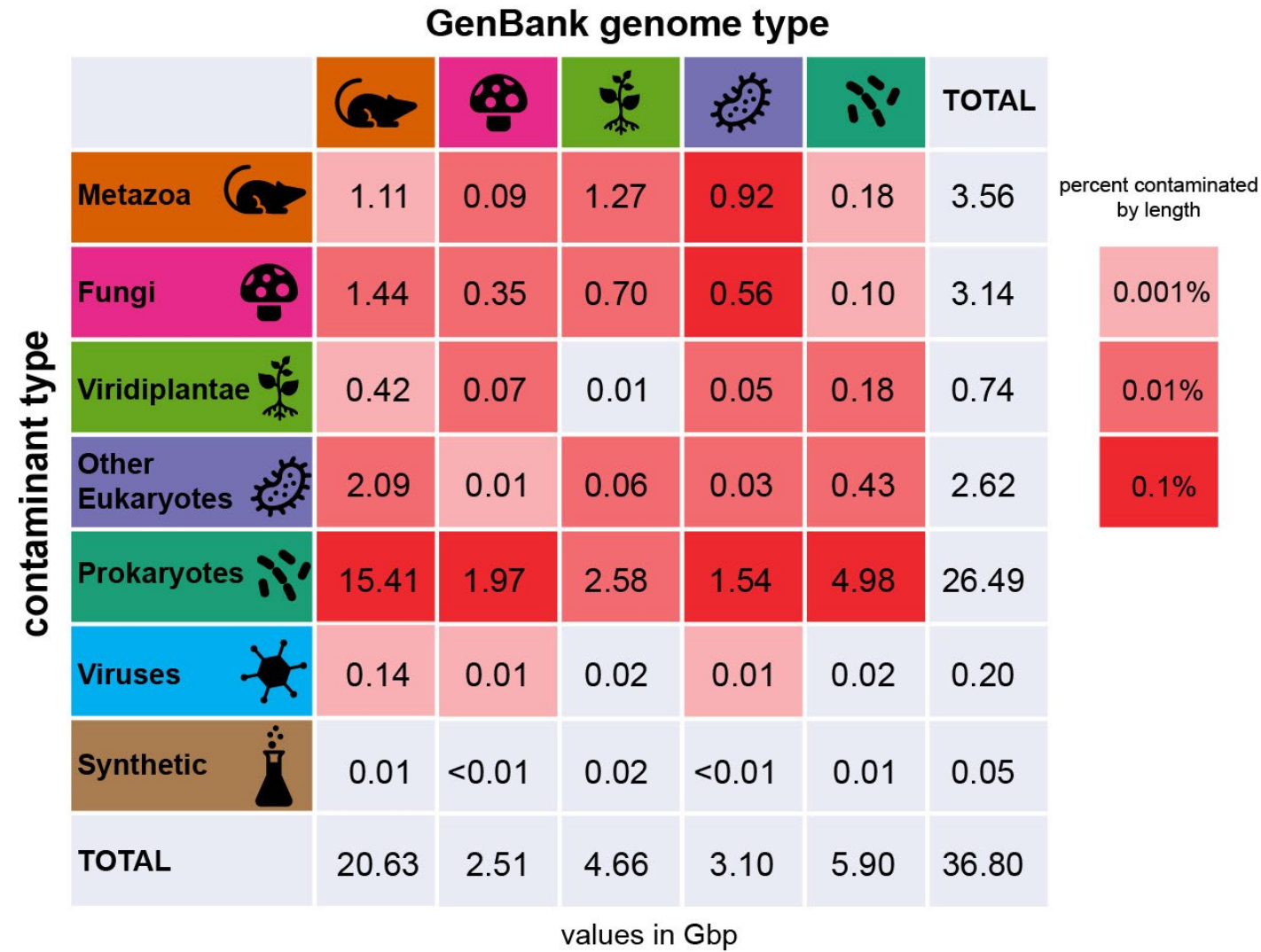
<https://github.com/ncbi/fcs>

<https://usegalaxy.org>



FCS-GX identifies contamination across the tree of life

- 36.8 Gbp of suspected contamination
- 0.16% bases
 - 1.30% sequences
 - Nearly 3 million annotated proteins
 - Half from 161 assemblies



EGAP: Eukaryotic Genome Annotation Pipeline

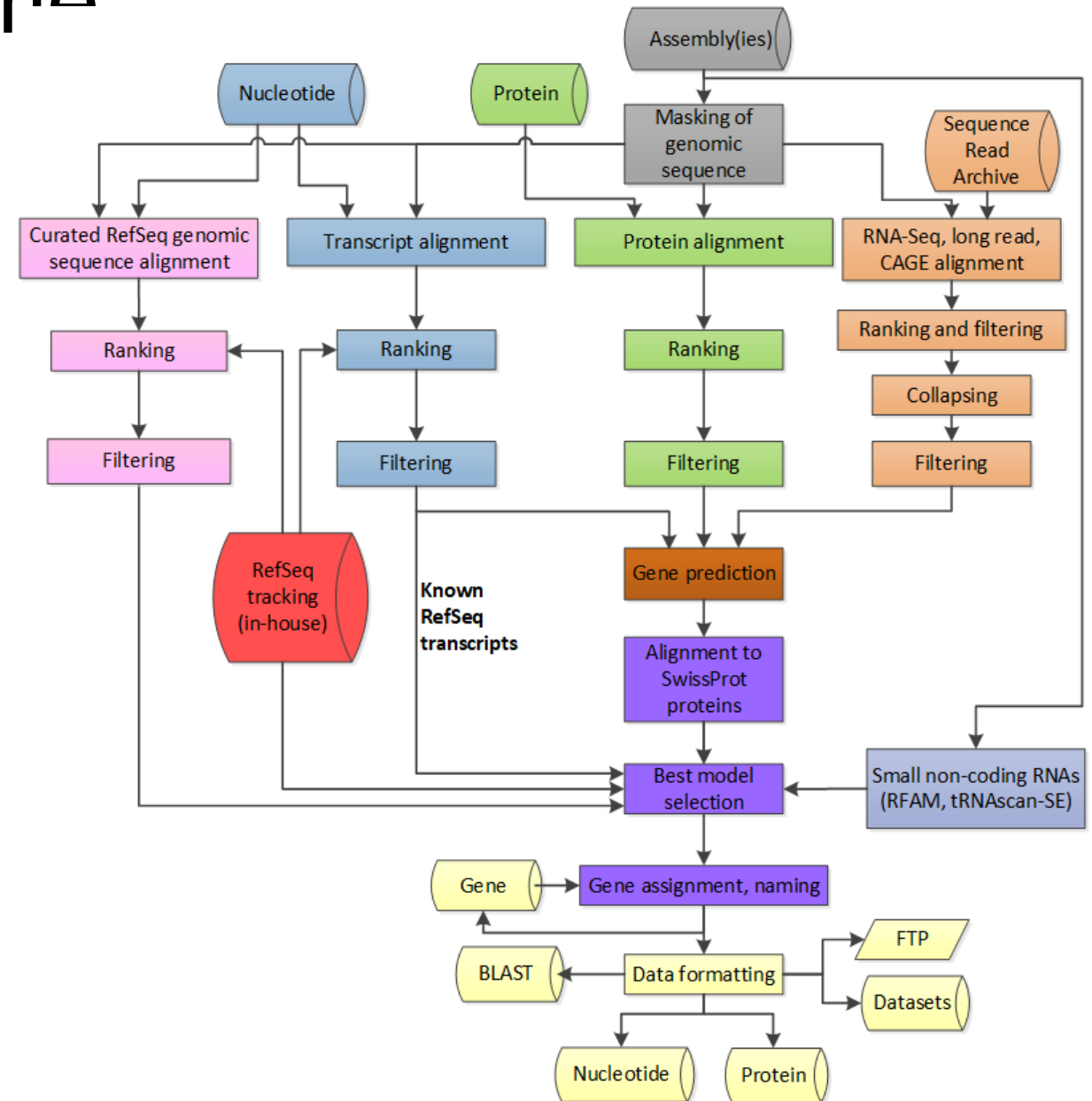
Used by NCBI to annotate ~1100 species

Evidence used for gene prediction:

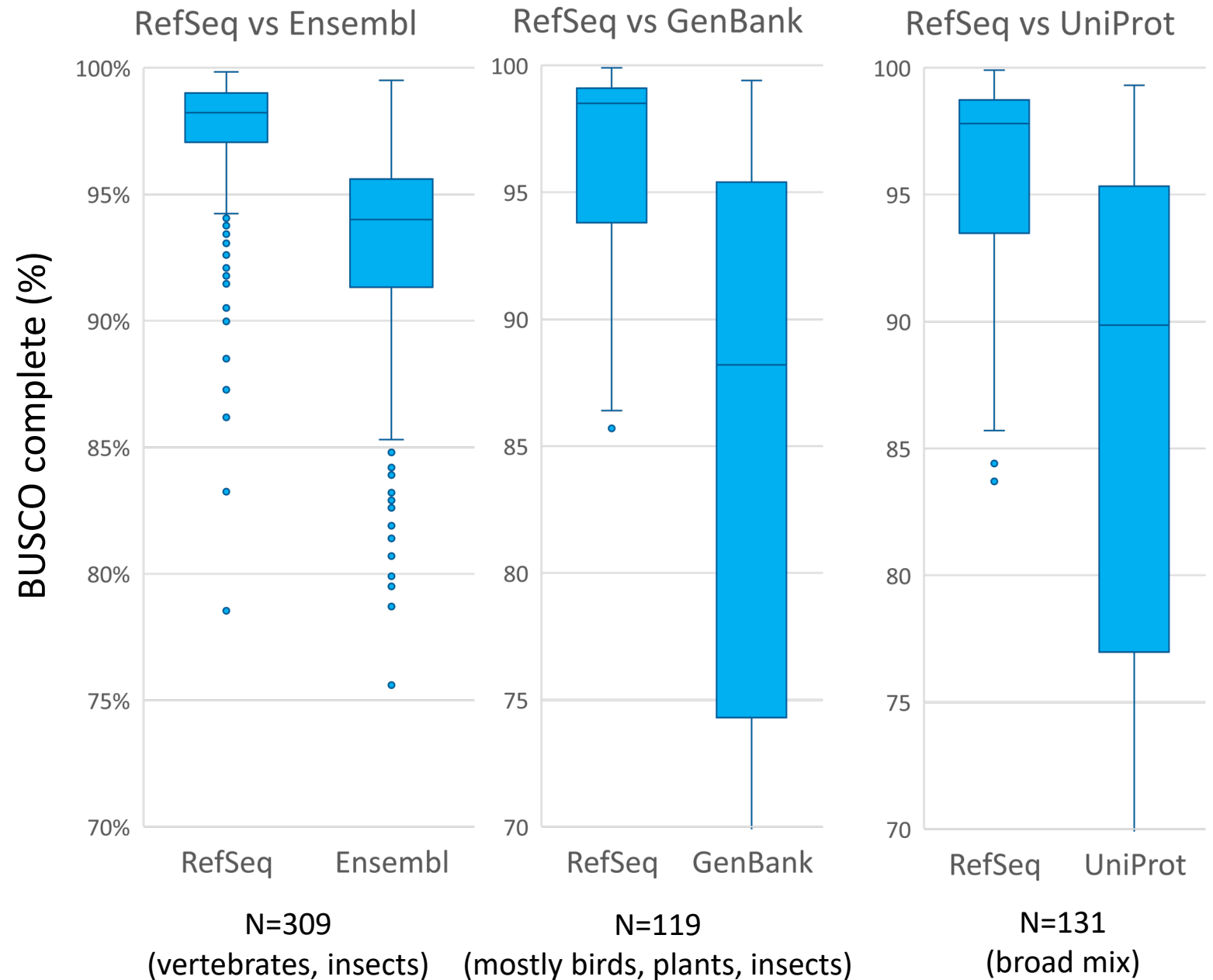
- ✓ ESTs
- ✓ cDNAs
- ✓ Same and cross-species proteins
- ✓ RNA-Seq
- ✓ PacBio IsoSeq, ONT transcriptomes
- ✓ CAGE
- ✓ PhyloCSF (for internal review)

Outputs:

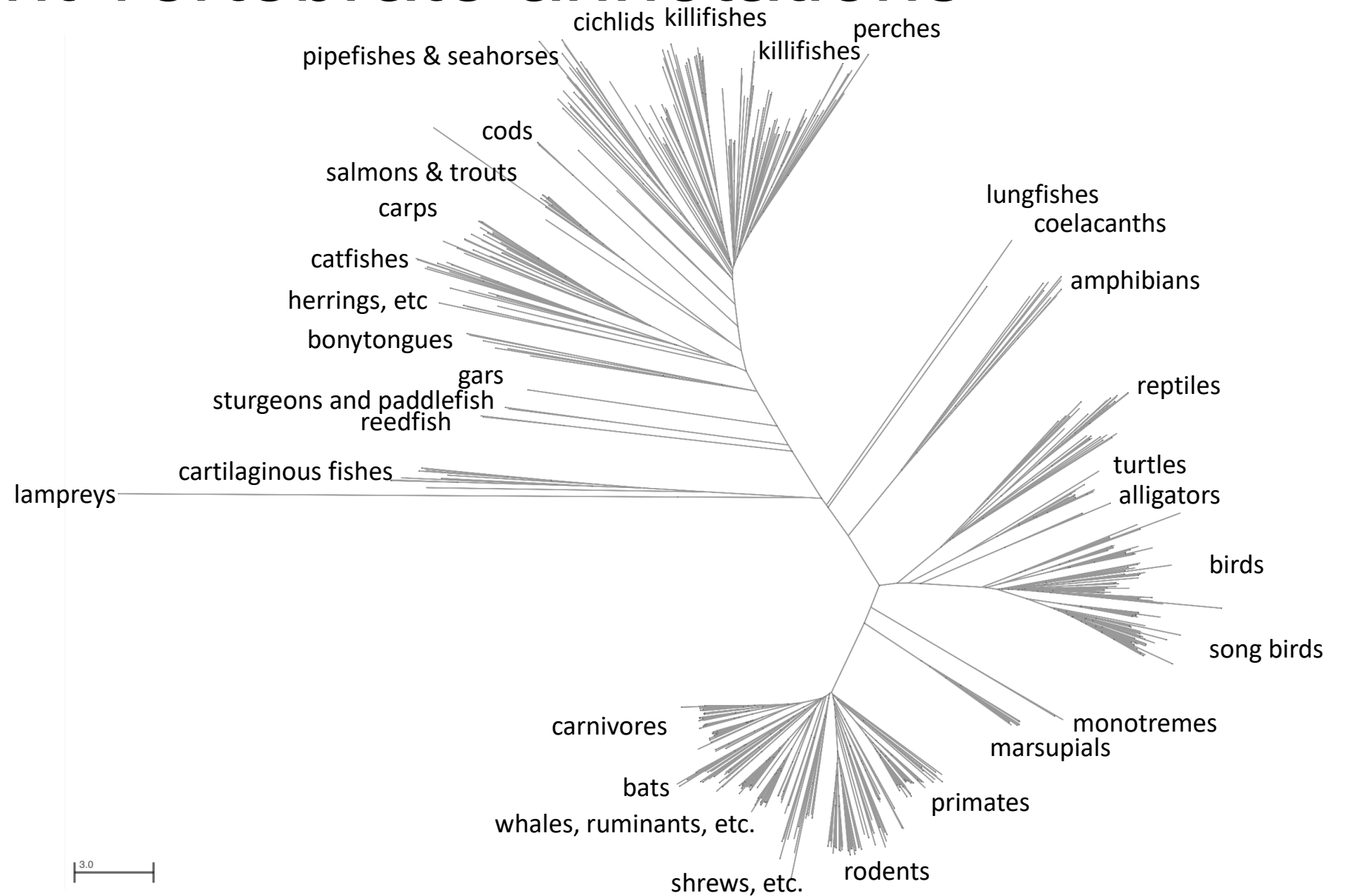
- ✓ Protein-coding genes
- ✓ Non-coding genes and pseudogenes
- ✓ Orthologs
- ✓ Gene Ontology
- ✓ Expression



More complete BUSCO models found in RefSeq annotations



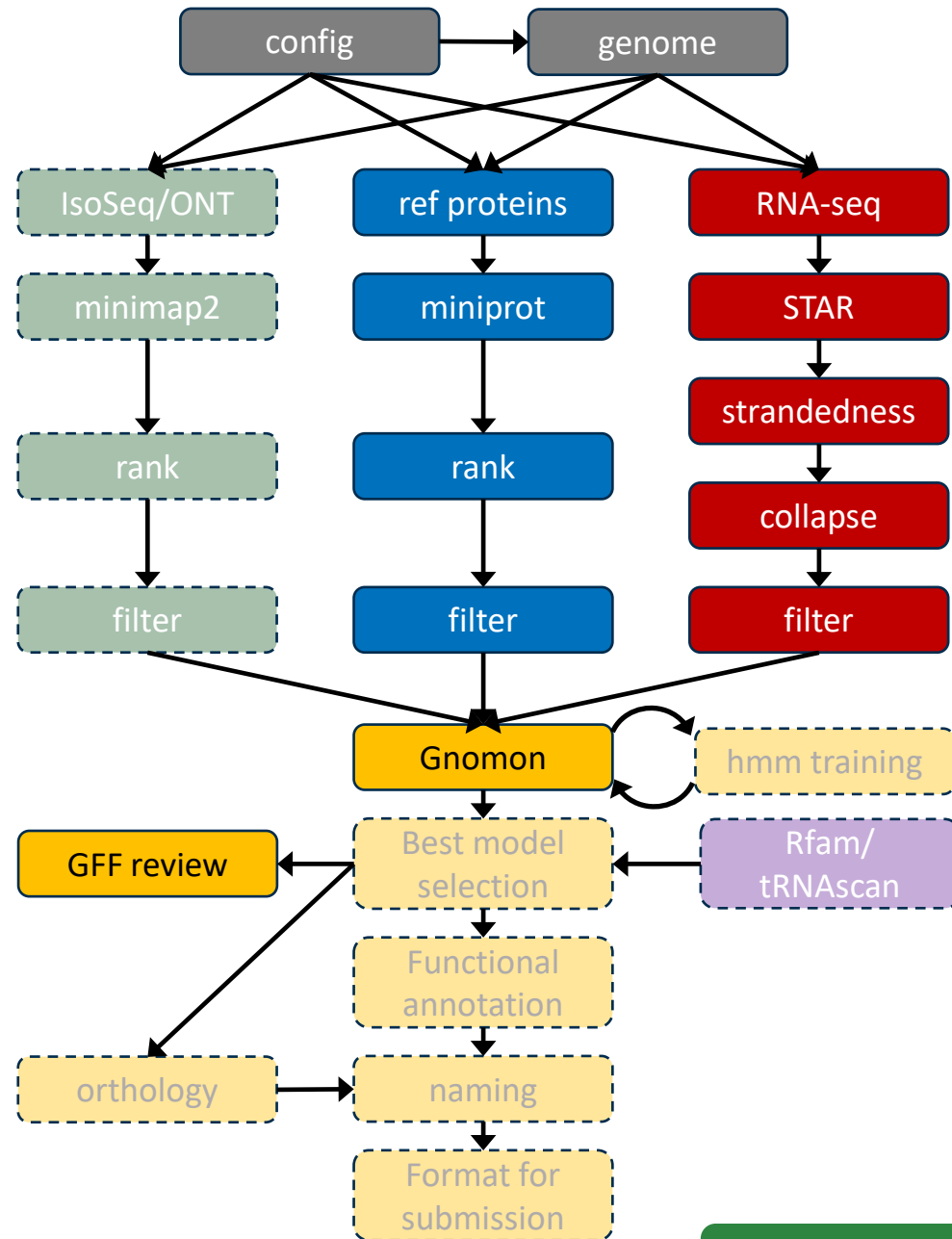
>580 current vertebrate annotations



Public EGAPx

Goals:

- NCBI EGAP that you can run yourself
- Nextflow pipeline available using Docker container
- Produce submission-ready annotation
- Work with public or private data
- Cloud-ready



Expanding annotation within a species

- NCBI RefSeq
 - Key organisms based on users or taxonomic placement
 - Focus on “reference” genomes (typically 1 per species)
 - Aim for highest quality annotation
- Additional assemblies
 - Annotate through EGAPx or LiftOff-type approach
 - Promote submission *with annotation* into GenBank archive
- Future support for pan-genomes combining the above

How Do I Get Involved?



Reach out to us at
cgr@nlm.nih.gov



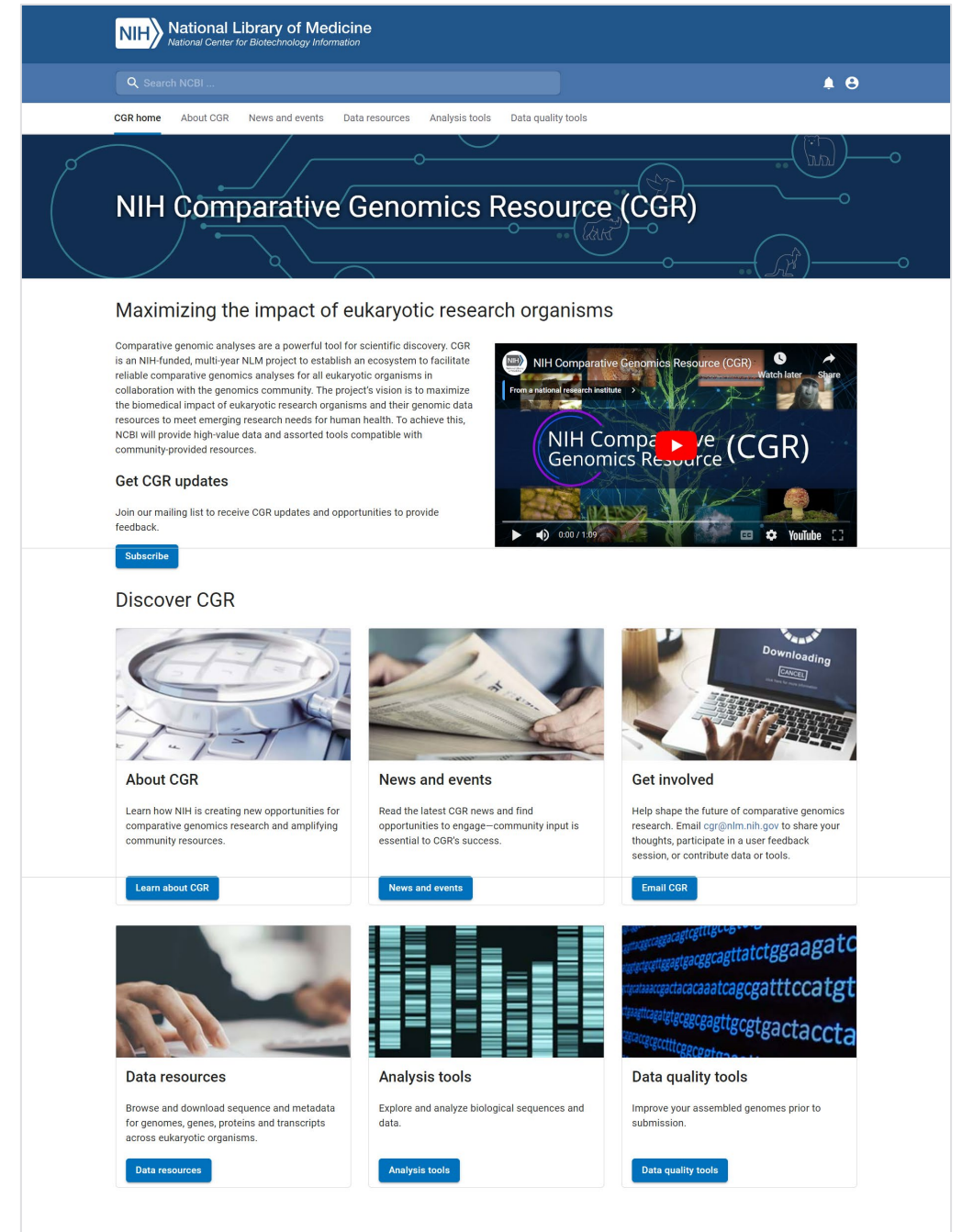
Visit the CGR website
ncbi.nlm.nih.gov/cgr
and click the yellow
Feedback button on the
bottom right of the page



Sign up for our
mailing list -
[bit.ly/Subscribe CGR](https://bit.ly/Subscribe_CGR)



Look out for future
meetings, workshops,
webinars, surveys, small
group sessions, user
testing, and interviews to
inform the development
process



The screenshot shows the NIH Comparative Genomics Resource (CGR) website. The header includes the NIH logo and navigation links: CGR home, About CGR, News and events, Data resources, Analysis tools, and Data quality tools. A search bar is also present. The main banner features the title "NIH Comparative Genomics Resource (CGR)" and a video player. Below the banner, there is a section titled "Maximizing the impact of eukaryotic research organisms" with a paragraph of text and a "Get CGR updates" button. Further down, a "Discover CGR" section contains six cards: "About CGR", "News and events", "Get involved", "Data resources", "Analysis tools", and "Data quality tools". Each card has a brief description and a corresponding action button.

NIH Comparative Genomics Resource (CGR)

Maximizing the impact of eukaryotic research organisms

Comparative genomic analyses are a powerful tool for scientific discovery. CGR is an NIH-funded, multi-year NLM project to establish an ecosystem to facilitate reliable comparative genomics analyses for all eukaryotic organisms in collaboration with the genomics community. The project's vision is to maximize the biomedical impact of eukaryotic research organisms and their genomic data resources to meet emerging research needs for human health. To achieve this, NCBI will provide high-value data and assorted tools compatible with community-provided resources.

Get CGR updates

Join our mailing list to receive CGR updates and opportunities to provide feedback.

[Subscribe](#)

Discover CGR

- About CGR**
Learn how NIH is creating new opportunities for comparative genomics research and amplifying community resources.
[Learn about CGR](#)
- News and events**
Read the latest CGR news and find opportunities to engage—community input is essential to CGR's success.
[News and events](#)
- Get involved**
Help shape the future of comparative genomics research. Email cgr@nlm.nih.gov to share your thoughts, participate in a user feedback session, or contribute data or tools.
[Email CGR](#)
- Data resources**
Browse and download sequence and metadata for genomes, genes, proteins and transcripts across eukaryotic organisms.
[Data resources](#)
- Analysis tools**
Explore and analyze biological sequences and data.
[Analysis tools](#)
- Data quality tools**
Improve your assembled genomes prior to submission.
[Data quality tools](#)

Thank you

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

CGR product team leads & admin

Terence Murphy
Nuala O’Leary
Sanjida Rangwala
Tom Madden
Aron Marchler-Bauer
Anne Ketter
Katya Sukharnikov

NCBI Leadership

Francoise Thibaud-Nissen
Valerie Schnieder
Kim Pruitt
Steve Sherry

Developers, subject matter experts and support

Kelsey Aadland	Emily Davis	Avi Kimchi	Dong-Ha Oh	Mirian Tsuchiya
Victor Ananiev	Olga Ermolaeva	Vamsi Kodali	Marina Omelchenko	Eric Tvedte
Preye Akuiyibo	Vladislav Evgeniev	Raymond Koehler	Yan Raytselis	Joel Virothaisakun
Ray Anderson	Robert Falk	Christopher Lanczycki	Dmitry Rudnev	Lukas Wagner
Alex Astashyn	Amelia Fong	Vadim Lotov	Victor Sapojnikov	Craig Wallin
Andrea Asztalos	Steven Gaudaen	Shennan Lu	Robert Smith	Jiyao Wang
Hena Bajwa	Marc Gwadz	Tom Madej	Alexandre Souvorov	Mingzhang Yang
Greg Boratyn	Vichet Hem	Gabi Marchler	Deacon Sweeney	Jian Ye
Evgeny Borodin	Wratko Hlavina	Patrick Masterson	Pooja Strobe	Irena Zaretskyna
Nathan Bouk	Jinna Hoffman	Scott McGinnis	Pape Sylla	Dachuane Zhang
Christiam Camacho	Brad Holmes	Yuri Merezuk	Jovany Tinne	
Eric Cox	Victor Joukov	Eyal Mozes	John Torcivia	

Contact us info@ncbi.nlm.nih.gov

Watch NCBI News for updates!

<http://www.ncbi.nlm.nih.gov/news/>
<https://www.youtube.com/user/NCBINLM>